Symmetry Breaking and the Emergence of Path-dependence

Abstract

Path-dependence offers a promising way of understanding the role historicity plays in explanation, namely, how the past states of a process can matter in the explanation of a given outcome. The two main existing accounts of path-dependence have sought to present it either in terms of dynamic landscapes or branching trees. However, the notions of landscape and tree both have serious limitations and have been criticized. The framework of *causal networks* is both more fundamental and more general that that of landscapes and trees. Within this framework, I propose that historicity in networks should be understood as *symmetry breaking*. History matters when an asymmetric bias towards an outcome emerges in a causal network. This permits a quantitative measure for how path-dependence can occur in degrees, and offers suggestive insights into how historicity is intertwined both with causal structure and complexity.

Keywords: Path-dependence - explanation - historicity - symmetry - causal networks - mutual information

Introduction

In many complex systems the past matters in explaining which outcome eventually obtains. This gives many processes in the special sciences, from chemical reactions to the evolution of political institutions, a seemingly irreducible historical character. The adoption of the QWERTY keyboard is often taken as a paradigm case of this phenomenon (David 1985). Originally, the QWERTY layout was designed in order to prevent typewriters from jamming; however, it subsequently became entrenched, even though the typewriter itself became obsolete, and even though there are more efficient ways of organizing an English-language keyboard for a computer. Thus, to explain why the present state of keyboards is as it is, one needs to integrate information about past states.

Perhaps the most prominent analysis of historicity has been in terms of *path-dependence*. In the broadest sense, path-dependence merely implies that the path followed by a system is explanatorily relevant for its final outcome. In this sense, the term is simply another way of saying 'past states matter'. However, once this 'explanatory relevance' is given a more precise characterization, narrower and more technical accounts of path-dependence emerge. Such accounts were originally proposed in economics and the social sciences (Arthur 1994, Pierson 2004), but more recently the issue has been receiving increasing attention from philosophers of science (Szathmáry 2006; Ereshevsky 2012; Desjardins 2011a, 2011b, 2015).

In the philosophy of science literature, the notion of path-dependence is closely related to two other ways of understanding historicity: 'sensitivity to initial conditions' (Ben-Yenahem 1997, Powell 2012) and 'contingency' (Beatty 2006, Desjardins and Beatty 2009). In this paper, these approaches will be integrated into the account of path-dependence (as also done by Desjardins 2011a) and therefore will not be explicitly discussed. After all, path-dependence in the broad sense is more or less synonymous with historicity to begin with, so these other approaches are best not seen as rival approaches, but as highlighting different aspects of path-dependence.

Among the more technical accounts of path-dependence, two classes of model can be discerned. The first characterizes path-dependence as occurring when a system could possibly evolve towards one of multiple local stable equilibria or *attractor states* (*e.g.* Bassanini and Dosi 1999, Pierson 2004, Szathmary 2006). In this way certain key aspects of path-dependence, such as nonlinearity and sensitivity to initial conditions, can be modeled as an evolution on what I call an *attractor landscape* with multiple attractor states. Such a model, like the adaptive landscape metaphor in evolutionary biology, has serious limitations, the main one being that in systems with high

dimensionality the topology of associated landscapes tend to be ridged and holey. As we will see, this means that the dynamics of such systems cannot be modeled as simply maximizing some scalar variable, and this precludes a general formulation of path-dependence in terms of landscape topology (see Gavrilets 2004, Kaplan 2008).

The second broad class of model has represented path-dependent processes as a *branching causal tree* (Desjardins 2011a). However, branching trees also have limitations when the causal structure becomes too complex, in particular when there are multiple possible initial states, or when there is a significant number of non-tree events, or 'reticulations', where branches converge (Moret et al. 2004). For example, an area where the tree metaphor has received significant criticism is phylogenetics, where phenomena such as hybrid speciation or lateral gene transfer cannot be captured in tree models.

To address these limitations, the first purpose of this paper is to introduce the notion of *causal networks* in some formal detail, and show how they are generalizations of both landscapes and branching trees. Network models are already well established in the causal modeling literature (following Pearl 2000) and in phylogenetics (e.g. Moret et al. 2004, Velasco and Sober 2010), but in the literature on path-dependence they have been underutilized. Causal networks allow complex causal relations to be represented when both tree or landscape metaphors fail.

The second, and main, purpose of the paper is to formulate a criterion of path-dependence that fits naturally within a causal-network framework. Borrowing from the notion of symmetry in physics, where it is used to characterize spatial configurations or dynamical equations, I will propose to extend the application of the notion of symmetry to the *space of possible causal paths*. The basic idea is that in ahistorical explanation there is a symmetry (interchangeability) between the biases towards the explanandum among all possible states at any given time. Path-dependence arises in a causal network when some past state has a different probabilistic bias towards the explanandum than the other contemporaneous past states — or in other words, when the symmetry is broken. Thus symmetry is potentially a powerful tool to characterize the path-dependence of a process without unnecessarily simplifying its causal structure to fit the mould of either a tree or a landscape.

In a final section I will outline how path-dependence can be quantified, as one unresolved challenge remaining is how precisely some processes can be 'more' path-dependent than others. I propose adopting the measure of 'mutual information', which is an information-theoretic concept used to quantify the amount of information one variable contributes about another. This measure is conceptually continuous with the symmetry formulation of path-dependence, and suggests one way in which the properties of path-

dependence can be studied formally.

I. Aspects of Path Dependence

In this first section I begin by selecting some salient properties of path-dependent explanations of processes[1] in order to develop some intuitions concerning the phenomenon. I will loosely group these properties according to whether they are future-oriented or past-oriented.

Among the past-oriented aspects, a key distinction is that between information-preserving and information-destroying processes (Sober 1983, 1988; Desjardins 2011). The latter is exemplified by what happens when a marble is released from the rim of a bowl: the marble will roll down and rock back and forth until it comes to a stop in the middle of the bowl. Given information about the final state alone, it is impossible to reconstruct its initial state, since no matter where precisely on the rim the marble was released, it would invariably have come to rest at the middle point. This is an example of path-independence, as the precise path followed by the marble makes no difference to the final state. In other words, the past of the marble is 'erased' and does not matter for the explanandum.

One of the most basic path-dependent processes is movement with friction. If one slides a block of wood from point A to point B, then it matters whether the shortest path between the points is chosen, or some more indirect route. In the latter case, more heat will be generated due to friction between the block of wood and the surface. Thus, some information about the past (*i.e.* the length of the path followed, or the speed with which the block was pushed) is preserved in the final state.

In general, most real processes have both information-preserving and information-destroying aspects, and in this way can only be said to be partially path-dependent. For example, the morphology of the whale shows remarkable similarities with the morphology of fish, yet there are significant differences as well. Some information about the past is destroyed due to the convergent evolution towards the streamlined morphology. However, a whale has lungs instead of gills, and its fins are exapted from fingers, and thus some information about its land-based past is preserved.

A second group of properties of path-dependence concerns how the past makes a counterfactual difference for the present: if the past were differ-

[1] In this paper, I take path-dependence to be a property of an *explanation* or a *representation* of a process, not a property of the process itself. A process is not identical with the representation of it, and epistemological problems arise because of this disjunction; however, this will not be a concern for the purposes of this paper. In interests of brevity, I will often refer to representations of processes simply as 'processes'.

ent, the outcome would also be different. For example, if humanity had skipped the technology of typewriters, and gone straight to computers, there would likely be no QWERTY keyboard. The phenomenon of *sensitivity to initial conditions* — how a small change in initial conditions can lead to a large change in outcomes — concerns this aspect of path-dependence (Ben-Yenahem 1997, Powell 2012). An example is the nonlinearity of the weather, so that, so to speak, a butterfly can flap its wings in Paris and cause a storm in New York. The outcome could not have occurred if the past were different.

The underlying notion here is the contingency of the explanandum, where 'contingency' refers not to the modal structure of the explanandum (*i.e.* whether or not it is true in all possible worlds), but to the structure of its causal history. The outcome is contingent if, given what we know about its causal antecedents, it could not have occurred.[2]

A helpful distinction here lies between causal-dependence contingency and unpredictability contingency (Beatty 2006). Causal-dependence contingency refers to the counterfactual dependence of the outcome on some prior state. Thus A is 'contingent upon' B if and only if, were B not present, A would not obtain. Causal-dependence contingency is thus a very broad notion, and also covers deterministic processes where there is dependence on initial conditions, such as the Newtonian dynamics of individual particles.

Unpredictability contingency refers more specifically to indeterminism in a process, or at least, a *modeled* indeterminism in the explanatory structure.[3] It is insufficient to know the prior states in order to predict the outcome state. Beatty describes this as 'contingency *per se*' (2006, 38-40), thus indicating that contingency can also be used as a one-place predicate attaching to an explanandum (Beatty 2006, 38-40).

These two notions of contingency capture two different senses in which the outcome could not have occurred, given the initial state. Unpredictability contingency is not necessary for path-dependence, as some perfectly predictable processes (*e.g.* Newtonian dynamics) depend on initial conditions, and thus the outcome is dependent on which causal path had been taken.[4] This has led some to refer to dependence on initial conditions as 'weak' path-dependence, and the cases where the process depends on multiple past states

[2]This is partially why historical explanations do not fit the mould of deductive (or even inductive) explanations. The explanandum cannot be deduced from a general principle, or inductively inferred with high probability, but maintains some degree of 'contingency'.

[3]Many processes in statistical physics and the special sciences are modelled as probabilistic, even though the underlying causal processes may be deterministic. See also footnote 1.

[4]We will see later on that unpredictability contingency is not sufficient either: some probabilistic processes are ahistorical.

as 'strong' path-dependence (Ereshevsky 2012).

These two orientations, future-directed and past-directed, are in no way mutually exclusive, and most real examples involve both perspectives. Consider the phenomenon of *positive feedback*, where the system is initially balanced between two basins of separate attractors, and where any initial fluctuation will snowball and result in a large, self-reinforcing change. A classic example of this is the emergence of the VCR videocassette technology (Arthur 1994). Initially, the videocassette market was precariously poised between two competing technologies, VCR and Beta; however, a slightly greater adoption of the VCR technology by consumers led to it becoming more widely available in video outlets, in turn precipitating further adoption by consumers. Thus VCR came to dominate the market. In this example, there is unpredictability contingency (the initial greater adoption was purely contingent) and sensitivity to initial conditions, but also some information-preservation, as given the outcome of VCR dominance, we can extract some information about some of the past states.

Another interesting combination of both orientations occurs in instances where the initial state neither snowballs nor is erased, but where it simply *constrains* future evolution. For example, developmental mechanisms, such as the processes determined by the *Hox* genes, constrain possible body-plans and thus the adaptations that are possible (Young and Hallgrímsson 2005). There is a counterfactual dependence in the sense that past states (like a certain configuration of the *Hox* genes) *preclude* some possible future states. When the past constrains the outcome to the extent that only a single outcome becomes possible, the phenomenon is known as 'entrenchment' or the 'lock-in effect'. There is also some information-preservation here, as it is possible to reconstruct the past to a certain extent.

II. Three challenges

With these phenomena in mind, three challenges face any account that attempts to uncover the more formal structure of path-dependence. The first is to account for how path-dependence is a matter of **degree**. While some measures have been intuitively suggested in the literature (*e.g.* Desjardins 2011a), a more rigorously developed account is lacking. This is partly due to the fact that the literature is relatively new. However, perhaps it is partly due to some confusion about two ways in which 'degree' can be understood.

The first way is when the past matters at multiple moments instead of a single instant. Thus, insofar the evolution of the whale is represented as depending on at least two moments (the transition from fish to land-based animal, and from land-based to aquatic reptile) instead of just one in

the case of VCR history (the instant when VCR happened to become more frequent than Beta), the evolution of the whale can be considered more path-dependent than that of the VCR. The outcome state gives more 'information' about the past in the first case than in the second.

The second way the past matters 'more' is when a difference in the past leads to a 'greater' difference in outcome. Thus the past matters 'more' when a butterfly flapping its wings leads to a hurricane than when the effects of the flapping are rapidly dissipated. Note that these two types of degree are not necessarily equivalent: for example, if the hurricane happens to be some attractor state, in such a way that *any* small disturbance would lead to the hurricane, the fact of a butterfly flapping its wings is not very interesting to explain why the hurricane occurred. More informative are the background conditions (pressure, temperature differentials) that had been forming; what actually triggered the hurricane relatively unimportant. In this way, the second measure of path-dependence does not concern how informative the past is for explaining the present, but concerns the 'distance' between possible outcome states.

In this paper I will leave this second sense of degree aside, mainly because the information-focused sense of degree is more fundamental and leads to interesting connections with information theory. However, another reason is that formalizing this second sense of degree would not be worthwhile for the purposes of this paper. Allow me to briefly sketch why. The distance-focused degree can either refer nonlinearity or discontinuity.[5] If one takes it to be discontinuity, it is a discrete property of a process and hence not a good candidate for a gradualist degree of path-dependence. If one interprets it as nonlinearity, then one would need to detail what it means for one outcome to be 'very different' from another. Which metric is one to use on, for example, the space of possible videocassette technologies? It seems impossible to introduce any metric without relativizing it to explanatory interests. Thus, if taken as nonlinearity, this distance-based degree of path-dependence seems to mainly depend on what explanatory interests are at stake rather than on the nature of path-dependence.

Besides accounting for how path-dependence comes in degrees, a second challenge is that the evolution of a system may be path-dependent at only

[5]In brief, a function is linear when $f(x + y) = f(x) + f(y)$; thus when a function is nonlinear a slight change in input will lead to an effect that is not linearly proportionate, and could potentially be very large. When a function is discontinuous, some modifications of the input, no matter how slight, will lead to relatively large effects. If a process is nonlinear but continuous, small changes will still lead to small effects; however, in a discontinuous process, some changes, no matter how small, will lead to lead to large effects, even if the process is otherwise linear.

certain times, or only with regard to certain outcome states. Thus path-dependence seems to have different **scopes**, some more local, others more global. A third, final challenge concerns the way in which path-dependence seems to depend on the **grain of analysis** adopted to describe the process. For example, the evolution of the whale is path-dependent when one distinguishes between the two states 'fish' and 'marine mammal'; however, path-dependence disappears when the outcome state is more coarsely described (e.g. 'aquatic animal').[6] What counts as an adaptation or a constraint is to some extent dependent on the grain of analysis (see Wilkins and Godfrey-Smith 2009). In general, introducing a more fine-grained description of the explanandum seems to make it more path-dependent: an account of path-dependence should be able to integrate this fact.

III. Establishing a framework

In this section I will outline two different frameworks that have been proposed to systematize these insights about path-dependence: attractor landscapes and causal trees. The second framework has been explicitly developed in some detail in Desjardins (2011a, 2011b), while I draw the first framework from a number of different accounts (Szathmary 2006, Bassanini and Dosi 1999, Desjardins 2015). Both frameworks have significant (but interesting) limitations, and are best seen as limiting cases of *causal networks*.

1. Attractor landscapes

An attractor is an equilibrium set of states towards which the system evolves when it is in a given neighbourhood (the 'basin'), and once in the attractor state, the system will tend to return there if perturbed. Its usefulness as a concept primarily lies in allowing for some long-term predictability, even in dynamics that are nonlinear and chaotic. An attractor is *global* when its basin covers all of state space, or *local*, when the basin is a subset of state space. In what I call an 'attractor landscape', each state is assigned a scalar variable (on a two-dimensional landscape, this is the height), with the attractors being local maxima (or minima), and the system tending towards maximizing (or minimizing) the scalar variable. Examples of such landscapes are potential energy landscapes, where valleys in the landscape correspond to minimal-energy states, or adaptive landscapes, where the peaks represent states with maximal fitness.

[6] See also Figure 6.

Landscapes can be used to systematize some aspects of path-dependence, for example, the distinction between information-preserving and information-destroying processes. Reconstructability becomes impossible when the explanandum (the outcome state) is a global attractor state, because any possible initial state tends towards the attractor state. When there are multiple attractors present, the process is partially information-preserving, as one can extract some information about the past (namely, in which basin the system was initially located) from the outcome.

With this in mind, one could formulate path-dependence in terms of the following negative condition:

Definition (Path-dependence - attractor formulation)**.** *An explanation of an outcome is path-dependent if and only if that outcome is not explained as a* ***global attractor****.*

Note that 'global' is always defined relative to the state space under consideration. The middle of the bowl is a global attractor when the state space is confined to the positions of the marble on the hollow surface of the bowl; it (obviously) is no longer an attractor when the marble is placed next to the bowl. Thus, when an attractor is deemed global within the scope of the explanation, then which precise initial state obtains does not make any difference for the outcome, as the system will be in the attractor state. Conversely, when there is no global attractor, then there are at least two initial states that lead to different outcomes.

The accounts of Bassanini and Dosi (1999) and Szathmáry (2006) implicitly draw on this criterion. Szathmáry distinguishes between 'strong' and 'weak' path-dependence (not to be confused with Ereshevsky's distinction). Strong path dependence occurs when the process is irreversible and when there are multiple stable attractors. This is straightforwardly covered by the attractor formulation.

However, what Szathmáry calls weak path-dependence could seem problematic for this definition. An outcome may not be a global attractor, and yet have occurred path-independently in the weak sense, as long as the causal-dependence on initial conditions is 'effectively' eliminated as time goes to infinity.[7] This type of weak path-dependence will tend to occur in high-

[7]Effective elimination is what Bassanini and Dosi (1999:15) call asymptotic path-independence, which occurs when two possible trajectories come arbitrarily close within a finite time-span, and for an infinite number of times thereafter. (If the dynamics is Markovian, then this condition reduces to: two possible trajectories intersect in finite time, because once there is a single intersection, it is expected that the paths will overlap for all subsequent times.) If this condition is met, then the difference an initial condition makes on a subsequent history is eliminated in finite time. In this way, weak path-independence is a form of ergodicity.

dimensional state spaces, when the number of possible states is 'much' greater than the number of states actualized over the course of a system's history, so that the asymptotic convergence of possible trajectories will tend not to occur (Szathmáry 2006).

Nonetheless, weak path-dependence is also covered by the attractor formulation, because the asymptotic convergence that Bassanini and Dosi describe concerns the convergence of the *average* position of a trajectory. Even though the actual *instantaneous* positions of two possible trajectories will in general be very different at any given time, when a system is weakly path-dependent, the long-run average position converges to a global equilibrium state.[8] Thus, whether or not the past matters in weak path-dependence depends on what precisely the explanandum is: the average position over a long period of time (past does not matter), or the actual position at a specific time (past does matter). By contrast, strong path-dependence implies that both the instantaneous and the long-run average position converge to a single attractor state.

The landscape framework has serious limitations. An area where it has already received significant criticism is in its application to biological evolution in the short- to middle-term (*i.e.* adaptive landscapes).[9] One important criticism concerns how landscapes change when the dimensionality of state space is increased. Landscapes imply that a system can evolve smoothly to a neighbouring state; Gavrilets (2004) has shown how the topology of high-dimensional adaptive landscapes tends to consist of'ridges', 'rugged peaks' and 'holes' than of smooth hills. The likelihood increases of a (nearly) neutral network forming, and of the number of local peaks increasing (see Gavrilets 2004, 45-80). The absolute scalar difference (in this case, fitness) between any two states becomes increasingly meaningless for predicting whether one state will evolve into the other or not.

What this suggests is that according as one needs more variables to characterize a particular outcome, the less likely one will be able to analyze the occurrence of that outcome as some kind of optimum of a single scalar quantity (*e.g.* fitness). While the attractor formulation of path-dependence may remain true, it becomes increasingly empty, as simple global attractors tend to not occur at all in complex systems. Knowing the height of a peak is increasingly useless as the landscape is increasingly ridged and holey.

[8]Compare with Doeblin's theorem in the theory of Markov processes (*e.g.* Stroock 2005).

[9]Note that while some have argued in their defense that they are best used as an explanatory template, as a heuristic for hypothesis generation (Ruse 1996, Skipper 2004), others have questioned their adequacy even as metaphor (Kaplan and Pigliucci 2006, Kaplan 2008, Plutynski 2008).

In this way, attractor landscapes cannot represent many interesting path-dependent processes; they are best suited to represent convergent processes, or processes where there is a choice between multiple local attractors.[10] Causal trees are better suited for evolutions in high-dimensional spaces, where the probability that causal paths intersect is very low, and thus where every state actualized is unique.

2. Causal trees

In the following I will briefly outline a formal characterization of trees, and then (drawing on the work of Desjardins) consider how path-dependence can be formulated within this framework. I will try to show that this framework is in a sense the opposite of attractor landscapes: best suited for high-dimensional state spaces, but weak at representing convergent causal structures.

A **tree** is a causal graph rooted in a single point, from which branches split off but never join as one moves from past to present. The states of a tree form a partially ordered set of states, where every state has only a single immediate predecessor, but can have any number of successors. Thus not every pair of states can be connected by a forward-directed causal chain, even though every state in a tree is indirectly causally linked through a common ancestor.

A causal tree maps out the possible paths an individual entity can follow. If the system consists only of a single entity, only a single path will actually be followed; if the system is an ensemble of individual entities, there will be a distribution over the possible paths according to the probabilities of the paths. The branching events or *nodes*, which connect a single state to two or more possible descendant states, can be thought of as abstractions of *contingent events* with causal impact on the path of the system. For example, in macroevolutionary phylogenetic trees the nodes abstractly represent speciation events, where a given biological population diverges to two or more distinct species.

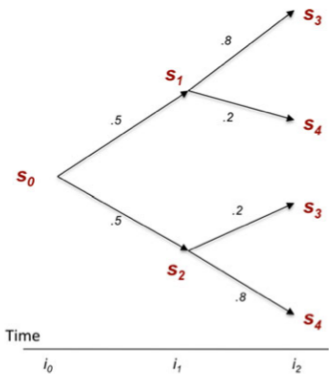

Figure 1: Path dependence in a causal tree (source: Desjardins 2011)

[10]Compare this with the analysis of conservative vector fields: if a dynamics can be represented as the gradient of a scalar, then it is path-independent.

With this in place, one can formulate path-dependence in the following way (adopted from Desjardins 2015)[11]:

Definition (Path-dependence - causal tree formulation)**.** *An explanation of an outcome is path-dependent if 1/ a given initial state branches off into at least two paths, 2/ these paths lead to at least two possible outcomes (with non-zero probability), and 3/ following different paths affects the probability of a given outcome state.*

This formulation of path-dependence captures some crucial properties, such as unpredictability contingency and causal-dependence contingency. It can also be used to capture the information-preserving aspect of path-dependence, and the way in which it can come in degrees (Desjardins 2011a). However, I would like to point to three limitations. The first and foremost is that, in contrast to attractor landscapes, causal trees cannot capture causal relationships where branches join. This is a problem for even the formulation of path-dependence, as path-dependence presupposes that there are alternative paths leading to the same outcome, and thus some convergence. This can be seen more clearly by redrawing Figure 1 so that the same states are represented by the same points; then the causal model becomes Figure 2, which is, strictly speaking, no longer a causal tree.

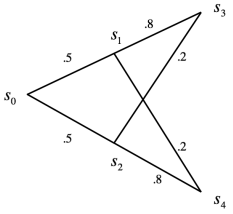

Figure 2: Path dependence in a causal network

Putting this problem aside (for example, by expanding the notion of tree to allow for some reticulations), it remains unclear how to analyze cases with multiple possible initial states. For example, in Figure 3, none of the paths leading to o_1 affect the probability of o_1 occurring, and thus the occurrence of o_1 does not seem to be path-dependent in the sense that its occurrence is not affected by the choice of path. Yet, there is a clear dependence on initial conditions, for if one knows that the initial state is s_0^*, the probability of o_1 occurring is .4, as opposed to .8, if s_0 were to be the initial state. The example in Figure 3 thus seems to involve some combination of path-dependence and path-independence that is not captured by the causal tree formulation.

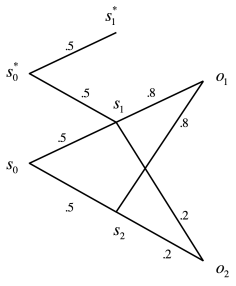

Figure 3: Path-dependent or not?

[11]For a more mathematical characterization, see Desjardins (2011a).

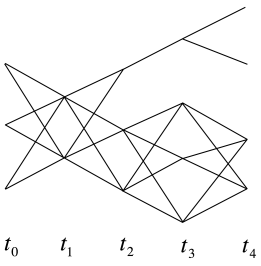

Figure 4: Path-dependent or not?

A second, related limitation is that the causal tree formulation concerns only whether the occurrence of an outcome is path-dependent, but it is unclear how it can be applied to a set or distribution of states, or how path-dependence is something that can change over time. In other words, the tree formulation does not seem to allow for different *scopes* of path-dependence. For example, in a more complex model such as Figure 4, there seem to be pockets of path-independence, even though the process may be globally path-dependent. The origin of this limitation lies less in the specific formulation of path-dependence, but rather in the causal tree framework itself; this is one important reason for representing causal relationships by causal networks (directed acyclical graphs).

Finally, it remains unclear how precisely the degree of path-dependence should be defined. Desjardins (2011a) suggests two types of metric that roughly correspond to those mentioned in section II. The first is the degree of divergence or convergence, where maximal divergence is maximal path-dependence and maximal convergence is maximal path-independence. The second is the degree of 'similarity' between outcomes: a causal tree is more path-dependent when the different outcomes are more dissimilar. However, it would be desirable to introduce a more precise, quantitative measure.[12] It is not clear, within a causal tree, what 'similarity' between outcomes could mean without introducing some independent scalar metric.

3. Causal networks

As done with causal trees, I will now construct causal networks with some more formal detail. Besides allowing for increased generality when describing path-dependence, there are two further advantages of doing this. The first is that it will become clear how a model can be *coarse-grained* to obtain either a causal tree or an attractor landscape, thus showing how the two frameworks are limiting cases of causal networks. The second is that it places path-dependence within the context of graph theory, to which the tools of information theory can be readily applied, and this will allow for a quantitative measure of path-dependence to be proposed.

A **causal network** is a directed, acyclical graph represented by the ordered pair (V, E), where V is the set of nodes and E is the set of edges

[12] Also, it can be shown that maximal divergence is, perhaps surprisingly, a case of maximal path-independence (see Figure 8).

connecting the nodes. In this paper, causal networks are taken to be formalizations of causal explanations, and hence certain nodes are of particular interest, namely the outcome states and the initial states. For this reason it will be useful to think of the ordered pair (V, E) as a 3-tuple (O, I, R) where O is the set of outcome states, I the set of initial states, and $R : O \to I$ a web of causal relations between initial and outcome states. The causal relations themselves may be productive or difference-making — the precise nature of causality will not be of concern here. In general, causal networks will contain intermediate states, between the sets of initial and outcome states. Letting these intermediate sets of states be represented by $O_i = I_{i+1}$, with $I = I_0$, $O = O_n$, the relation R can be decomposed in $n + 1$ instants: $R = R_0 \circ R_1 \circ \cdots \circ R_n$, where each $R_i : I_i \to O_i$ is a simple mapping relation.

Three basic causal patterns will be of interest. In a **parallel** structure, the outcome would not have obtained if a particular initial state had not been present. Thus, there is at most one initial state associated with a given outcome, and in this way the parallel structure corresponds to causal-dependence contingency. By contrast, in a **divergent** structure, multiple outcomes are associated with a single initial state. This means that, given the initial state, the descendant state cannot be predicted: this is unpredictability contingency. When the structure is neither parallel nor divergent, it is **convergent**, and this occurs when multiple initial states converge on a single outcome state. A path-dependent explanation, as actually used in scientific practise, is almost invariably a complex combination of these basic structures.

The **probability of an outcome** in a particular explanatory framework can be calculated by means of the probability distribution over initial states, and the probabilities of the different paths between an initial state and the outcome. By the law of total probability we get $P(o) = \sum_i P(i)P(o|i)$. Each conditional probability $P(o|i)$ can be written as $P(o|i) = \sum_p P(p_{io}) = \sum_p \prod_{i \leq j \leq o} \pi(j \to j + 1)$, where the p_{io} are the different paths connecting initial state i to outcome o, and where $\pi(j \to j + 1)$ represents the transition probability connecting two intermediary states. Thus the probability of an outcome is ultimately reducible to the initial probability distribution and the structure of the causal pathways leading up to the outcome.

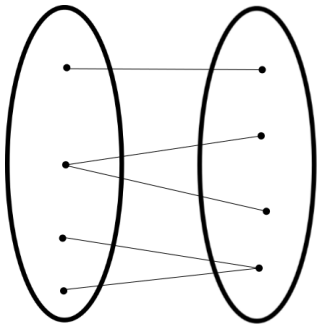

Figure 5: Parallel, divergent and convergent structures.

In general, the causal structure changes by fine-graining or coarse-graining the degree of analysis. Fine-graining can be thought of as introducing a

new variable to characterize the initial or outcome states, and in this way states that were previously identical become differentiated. More explicitly, a state may be characterized by n variables, $s = (x_1, x_2, ..., x_n)$, and one example of fine-graining is to introduce m new degrees of freedom, defining a new state $s' = (x_1, x_2, ..., x_n, ..., x_{n+m})$. So a single state s in n-dimensional space corresponds a m-dimensional set in the fine-grained (n+m)-dimensional space. In this way fine-graining can be thought of as a one to many mapping, where a single state is replaced by a set of states.

The inverse operation is coarse-graining, and this is done by means of an *equivalence relation* $\sim$, which allows one to express that multiple states are 'similar' in some way. The equivalence relation defines an equivalence class on the states, $O/\sim$, where all the states which are 'similar' are represented by a single state. One way this can be done is by abstraction, where certain degrees of freedom are dropped, so that only the other features of a state are considered.

This offers a first step in making sense of how path-dependence is sensitive to the grain of analysis adopted in a causal model. Taking the whale's evolution as an example, what is striking here is that there is both convergence towards a fish-like morphology, and a divergence in other respects (such as bone-structure or respiratory system). One way to analyze this is that, when the aquatic mammal state (AM) and fish state (F) are characterized by a single variable — their overall morphology — is that they are the same state; when the two states are characterized by additional variables (bone-structure, respiratory system, *etc.*), the two states are non-indentical. In the first case, the paths $F - M - AM$ and $F - F - F$ converge; in the second, fine-graining introduces path-dependence in the representation of the evolutionary process (figure 6).

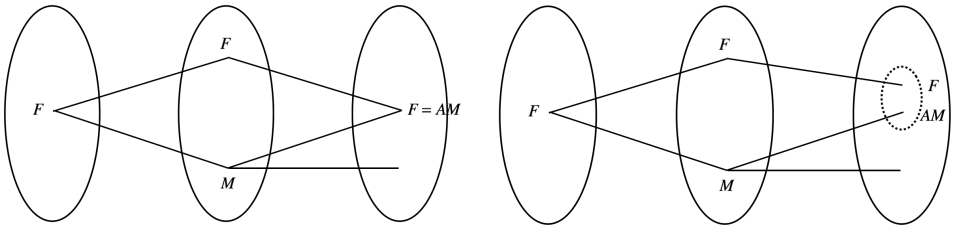

Figure 6: Two representations of the evolution of the whale. The right side representation is models the evolutionary process in detail, and is path-dependent. The left side one coarse-grains over the aquatic mammal state (AM) and fish state (F), and represents the evolution of AM as path-independent (or, at least, more path-indepedent: see later).

One can summarize the effects of the grain of analysis on causal structure by means of the following (a proof is provided in the appendix):

Theorem 1. *A coarse-graining of the explanandum makes an explanation increasingly convergent and a coarse-graining of the explanans makes an explanation increasingly divergent.*

This theorem gives some deeper insight into why attractor landscapes and causal trees are limited. In any attractor landscape, there is a countable number of privileged outcomes (attractor states), and each of these outcomes will have an associated subset of the initial states (the basin). When state space is described at a finer level of detail — *e.g.* when more variables are needed to adequately describe each state — the convergence of each basin on its respective attractor state will tend to decrease. A given attractor state will be disambiguated between two different states, each with its own basin. Ultimately, when the outcome states are described with sufficient detail, there will be no convergent structures any longer, only parallel structures, and the landscape metaphor disintegrates.

By contrast, the causal tree framework tends to be adequate as long as the number of possible states is much greater relative to the number of realized initial states, so that the probability of reticulations occurring is small. For example, this occurs when the dimensionality of the state space is relatively large. Taking the number of variables necessary to describe an entity to be a proxy for the complexity of that entity, this can also be formulated in terms of complexity. The dynamics of an individual, complex entity is likely to be path-dependent. By coarse-graining the state space (representing the complex entity abstractly) while keeping the number of initial states constant, the convergence of the network increases monotonically, and the 'tree-ness' of the network decreases. In this way a causal tree can be seen as the limiting case of a causal network when the state space is much larger than the set of initial states.

IV. The symmetry formulation

In this section, the main contribution of this paper, I will propose how the concept of *symmetry breaking* can be used to characterize path-dependence and historicity in causal networks. The motivation for this proposal comes from the two main ways symmetry is used in physics (see also Brading and Castellani 2007). The first, and most intuitive application of symmetries is to *properties* of a system, usually spatial configurations. A spatial configuration is symmetrical when it remains the same under some distance-preserving

permutation of the elements (reflections, inversions and rotations). For example, a snowflake has some rotational symmetries (its appearance is unchanged when you rotate it by a multiple of 30°), reflection symmetries and a point symmetry. Similarly, a liquid has a maximal spatial symmetry: no matter how one would rotate, invert or reflect it, it would look the same. Such symmetry is *broken* during the transition to a solid: a particular molecular structure arises which will typically only have a limited number of symmetries.

Symmetries are also applied to the *dynamics* of a system, *i.e.* the way in which two subsequent states of a system relate to each other. Thus, instead of transforming the physical elements of the system, the variables in the laws of motion are transformed, and a symmetry is said to be present when the laws of motion remain invariant. In other words, the transformation is a symmetry of the dynamics if the transformed variables are related to each other in the same way as the untransformed variables are. One well known example is the time symmetry of Newtonian dynamics: because the second law gives a relation between the force and the second time-derivative of position (*i.e.* the acceleration), it is invariant under the transformation $t \to -t$. Thus, if one were to see an animation of a group of interacting particles, one could not tell by Newtonian dynamics alone whether the animation was being played forwards or backwards. In thermodynamic phenomena this time symmetry is broken: heat flows from warm to cold (the entropy increases), but never from cold to warm. A rewinded heat flow does not obey the second law of thermodynamics.

Here I will apply symmetry to the causal paths between intermediary states and a particular outcome. A network will be symmetrical when the different intermediary states can be permuted without affecting the causal structure of the network. Just as the snowflake remains unaffected by rotations, path-independent causal networks remain unaffected by permutations of intermediary states. In itself, this basic idea is not much more than a reformulation of path-independence in the broad sense; however, it offers the resources to deal with some of the shortcomings of the tree and landscape frameworks.

1. Symmetry

More formally, let P_s be the probability distribution over the outcomes given that the system is in state s. One way to think of P_s is as the probabilities of the different possible outcomes as 'viewed from' s. The probability of any particular outcome o as viewed from s can be written as the sum of the

probabilities of the different possible paths between s and o:

$$P_s(o) = P(o|s) = \sum_p P(p_{os})$$

where the variable p_{os} represents the possible paths between o and s. When there is only a single initial state s_0, one can assign an unconditional probability to an outcome $P(o) = P_{s_0}(o)$. This is the case in causal trees; however, in a general causal network, there is no unique way of specifying the unconditional probability of an outcome.

Note that these probabilities need not imply any fundamental indeterminism. For example, in ecological systems of foraging rabbits, the dynamics of how rabbits move around may not be fundamentally indeterministic, and may be perfectly predictable if, for example, the position, visual cues and neural states of the rabbits are perfectly known. Yet, we may choose to ignore such details, and to characterize the state of a rabbit in terms of position only. This is obviously an underdeterminantion, and multiple outcomes will be possible given the same position. In this way, coarse-graining and even ignoring certain variables can give rise to probabilistic causal relations (see Strevens 2006, Matthen 2009). For purposes of this paper the precise nature of these probabilities need not concern us further, and we will treat them simply as given.

The notion of causal symmetry can be assigned different scopes, some more local, others more global:

Definition (localized to time and outcome)**.** *A causal network is* ***causally symmetric towards outcome o at time t*** *when the biases of any two states s and s' at time t towards o are equal: $P_s(o) = P_{s'}(o)$.*

This notion of symmetry is relevant for the question as to whether a particular instant in the past matters for a particular outcome. When the explanatory interest concerns the question whether *any* past state matters for a particular outcome, the following scope of symmetry is more appropriate:

Definition (localized to outcome)**.** *A causal network is* ***causally symmetric towards outcome o*** *when the biases of any two states s and s' towards an outcome o (at any time t) are equal: $P_s = P_{s'}$.*

This type of symmetry corresponds most closely to how path-dependence was formulated in the causal tree formulation, except that now an allowance is made for multiple possible initial states. Symmetry can also be localized to time alone:

Definition (localized to time)**.** *A causal network is* ***causally symmetric at time t*** *when the biases of any two states s and s′ at time t (towards any outcome o) are equal:* $P_s = P_{s'}$.

Finally, a properly 'global' notion of symmetry can be formulated, as to predicate path-dependence about an explanation as a whole, not just an outcome: a network can be said to be **causally symmetric** when it is causally symmetric at every time t (or equivalently, towards any outcome o). This concept of global symmetry entails the three local notions of symmetry, and the most localized notion of symmetry is implied by the three others.

The transformation group associated with global symmetry is the group of permutations of the intermediary states at any given time. Global symmetry arises when the conditional probability distribution over the outcomes remains invariant under permutation of the intermediary states at any given time (this includes the initial states).

Figure 7 illustrates how these four scopes of symmetry can diverge. First, the network is not globally symmetric, since, for example, $P(o_1|s_4) = 1/2 \neq 0 = P(o_1|s_5)$. Thus, in order to explain why o_1 occurred, it is relevant that s_4 and not s_5 occurred. However, the network is symmetric with regards to some outcomes at some particular times. For example, at t_5 the network is symmetric towards o_2 as $P(o_2|s_3) = P(o_2|s_4) = P(o_2|s_5) = 1/2$. It does not matter what state the system is in at t_5 to explain why o_2 occurred. Similarly, the network is symmetric towards o_3 at t_3.

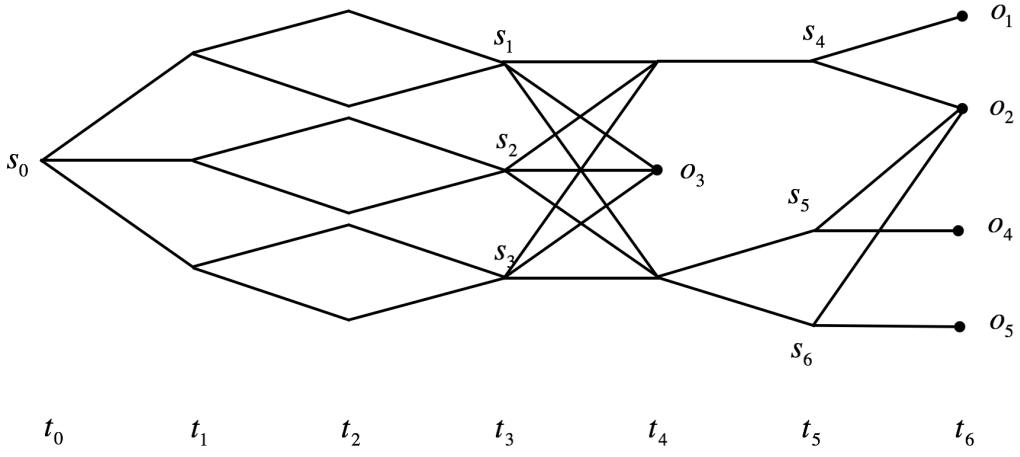

Figure 7: An illustration of how the different notions of symmetry come apart. Each state branches out in an equiprobable way.

Concerning the two other notions of symmetry, the network is symmmetric at t_3, as the biases of s_1, s_2 and s_3 are equal towards *any* of the outcomes

o_j. In an explanation of any outcome, it will not matter what state the system was in at t_3. Finally, the network is symmetric towards o_3. To explain why o_3 occurred, it will not be necessary to integrate *any* information about the past. Regardless of the path the system took, o_3 would have occurred with probability 1/3.

Deepening the parallel with spatial symmetries, causal symmetry can be given a geometric interpretation within a causal network. A network is symmetric at time t if every state at that instant t branches out to all descendant states in an identical way. Thus the branching pattern emitted by one state must be mirrored by all other possible states at that time. This basic pattern is represented in Figure 8, where the thickness of the lines is a measure for the probability of the different transitions. Some descendant states may be very improbable while others may be heavily biased; what matters is that the biases are symmetric across the different initial states. At a symmetric instant in the network, the different states can be exchanged and permuted without the causal structure being affected.

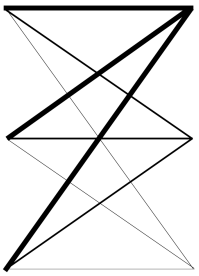

Figure 8: Fundamental pattern of symmetry.

This basic pattern of symmetry is both *maximally divergent* and *maximally convergent*. It is maximally divergent because each state branches out towards all possible descendant states; it is maximally convergent because each descendant state is converged upon by all possible predecessor states.

Anticipating the next section, where symmetry is linked with path-independence, this fact suggests that path-dependence is to be sought between the extremes of perfect predictability and perfect unpredictability. Both the perfectly predictable network — where all paths converge onto a single outcome — and the perfectly unpredictable network — where all states diverge maximally — are ahistorical. Path-dependence requires some degree of unpredictability, but maximal unpredictability contingency implies path-independence. This is a concrete result that precludes any simple identification of unpredictability contingency with historicity (*e.g.* Beatty).

An additional effect of the basic pattern of symmetry is one of *erasing history*. In Figure 7, the network up until t_3 could be replaced by any arbitrary causal network without any difference being made to which outcome obtains. This effect is encapsulated in the following result:

Theorem 2. *If a causal network is symmetrical at t, it is also symmetric at all prior instants. The bias of any state towards a given outcome is shared throughout the states at any given time, and is preserved over time.*

Thus a sufficient condition for global symmetry is that only the last causal transition is symmetrical, *i.e.* each direct parent of the outcome states branches out to all outcomes. Note that, given such a symmetrical structure, none of the intermediary states affects the outcome, and hence there is no history to erase, strictly speaking. History matters only to which intermediary states occur (even though these intermediary states are not the target of the explanation), and before the occurrence of the symmetrical pattern, it is possible to reconstruct the past. Once a symmetrical pattern occurs, such reconstruction is impossible.[13]

A concrete example that could be represented by such a causal structure is mass extinctions. To the extent that one can idealize mass extinctions as the random selection of certain phenotypes (without regard to fitness), it is impossible to reconstruct the distribution of phenotypes *before* the mass extinction given the distribution *after* the extinction.[14] Even though non-symmetric processes may have dominated up until the point of the mass extinction, once the mass extinction has taken place, the effect of these processes on history is wiped out.

2. Symmetry breaking

These different notions of symmetry are different ways in which the past does *not* matter, different ways in which the system is *in*dependent of the path taken. Path-dependence itself can be formulated as the *breaking* of symmetry, and thus has different scopes as well.

Definition (Path-dependence - symmetry formulation)**.** *A causal network is path-dependent relative to a certain scope if and only if the symmetry relative to that scope has been broken.*

In this way, a network may be globally path-dependent even though at certain times it may be path-independent, or even though certain outcomes may emerge in a path-independent way.

The attractor and causal tree formulations of path-dependence can be seen as special cases of this more general definition. If a causal network converges onto a global attractor, this means that any two states s and s' at any time t will lead to the outcome with probability 1: $P_s(o) = P'_s(o) = 1$. Conversely, if the outcome is not a global attractor, there is at least one

[13]Thus it is impossible for history to matter for an outcome some time in the past, but not at the penultimate stage (compare with Desjardins 2015).

[14]In this way, while mass extinctions introduce contingency into evolution (as famously emphasized by Gould 1989), to the extent that they make the reconstruction of the past more difficult, they actually remove some degree of historicity.

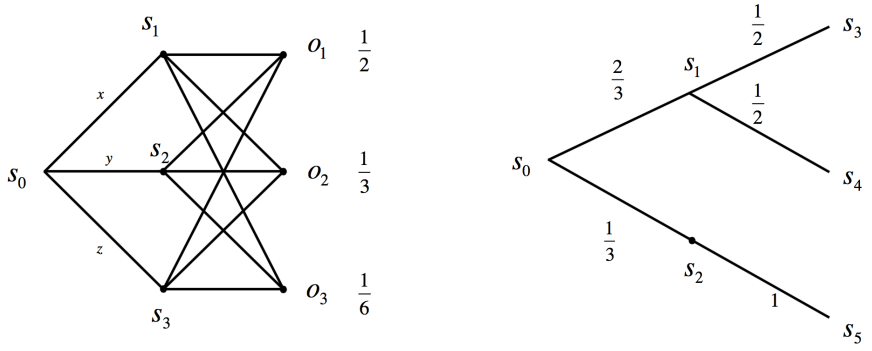

Figure 9: Unconditional probabilities of outcomes vs. path-dependence.

possible state that is not in the basin of that outcome. In this case there are at least two states s or s' that have a different bias towards the outcome at some time t: the symmetry towards o is broken.

In the tree-framework, path-dependence was limited to comparing possible paths leading to one of a number of possible outcomes. The requirements of the causal tree formulation of path-dependence – there must be multiple possible outcomes (*i.e.* so that convergence can only be partial), and that paths towards some outcome affect the probability of the outcome – are captured within the negation of symmetry (localized to time and outcome). These requirements can be deduced by the condition that at least two states on different paths have a different bias towards the outcome.

The significance of this definition may be further illustrated by pointing out what it does *not* entail. First, it does *not* entail that no outcome is probabilistically privileged. Some outcomes may be more likely than others, and yet the network is symmetrical; all outcomes may be equiprobable, and the network path-dependent (Figure 9). The unconditional probability of an outcome is irrelevant; what matters is whether the conditional probabilities are equal or not.

A second orthogonal distinction is between path-dependence and the probabilities of the paths. The occurrence of an outcome may be path-independent, even though some paths may be heavily biased. For example, on the left side of Figure 9, there are three possible paths towards o_1. Even if the system may be much more likely to pass by s_1 than the other intermediary states (*e.g.* $x = 0.98$ and $y = z = 0.01$), $P(o_1|s_i) = 1/2$ for each intermediary state s_i. History does *not* matter: it makes no difference whether the system takes the s_1 path or the s_2 path, in each case o_1 will obtain with probability $1/2$.

Thus, in a path-independent network it may be possible to reconstruct

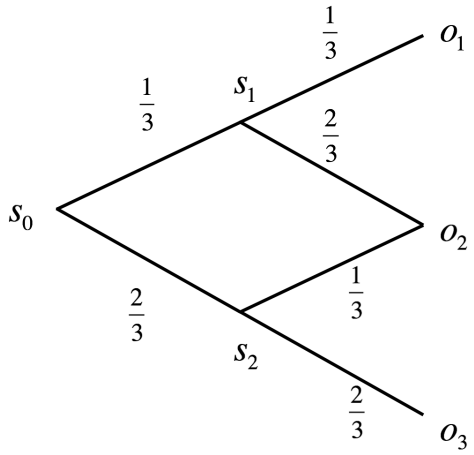

Figure 10: Symmetry towards o_2 is broken, even though all paths to o_2 equiprobable.

the past; conversely, retrodictability may be impossible in a path-dependent network. Such is the case in Figure 10, where the two possible paths towards o_2 are equiprobable, but yet where the symmetry is broken at the intermediate states since $P(o_2|s_1) = 2/3 \neq 1/3 = P(o_2|s_2)$.

There is no retrodictability here since, given that the system is in o_2, it is equiprobable that the system passed through s_1 as through s_2.[15] The relation between retrodictability and path-dependence will be taken up again in the final section, but since this result may seem puzzling here, one can illustrate it with an example. Say that s_1 represents 'financial crisis' and s_2 represents the avoidance of a financial crisis. The outcome state o_2 is a state of revolution. A financial crisis may be very improbable, but yet, once it occurs, revolution may be very likely. Conversely, a revolution may occur spontaneously with a very small likelihood. Even though these two paths may be equiprobable, if society actually underwent a financial crisis, any historian would integrate this information to explain the outcome.

3. Symmetry preservation and restoration

An additional advantage of the symmetry formulation is that it can distinguish between different scopes of path-dependence. Certain parts of a causal network may behave in a path-independent way, even though the network as a whole is path-dependent. The past may not matter in the causal explana-

[15] By Bayes' rule, $P(s_1|o_2) = \frac{P(o_2|s_1)P(s_1)}{P(o_2)} = \frac{2/3 \cdot 1/3}{4/9} = 1/2$.

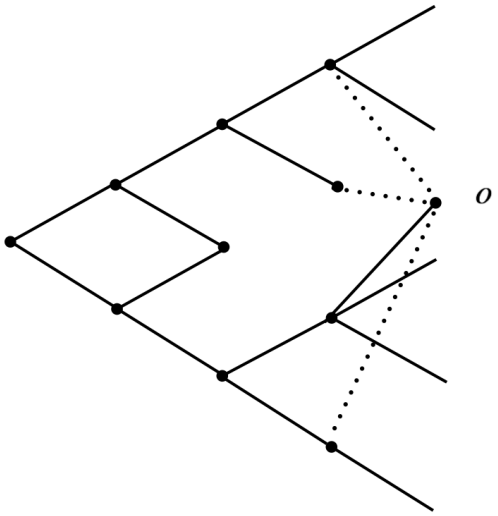

Figure 11: Weak global attractor: convergence and path-dependence.

tion of a particular outcome, but yet may matter in the explanation of the set of outcomes. Or, the evolution of the system may be path-independent until a certain moment in time, after which the causal network becomes path-dependent. Path-dependence (localized to time) can emerge at a particular instant in the causal network.

Two combinations are of particular interest: cases where symmetry towards a particular outcome is preserved, despite global symmetry being broken, and cases where global symmetry is restored for a subset of the causal network. An example of the first is represented in Figure 11. Here the outcome o is a global attractor in the sense that all possible initial states can evolve towards o, and the occurrence of o is path-independent as all prior states are equally biased towards o. Yet global symmetry is broken in the network as a whole.

Such a state o can be termed a *weak global attractor*: a state that remains a possibility with a fixed probability regardless of the path the system takes. When a weak global attractor is present in a network, a local symmetry is preserved, even though the global symmetry may be broken.

The second case of particular interest concerns states that branch out towards all possible descendant states in an equiprobable way. Evolvability would be a concrete example of this causal structure.[16] For example, in most mammals, forelimb and hindlimb are locked by certain developmental constraints in a 1:1 ratio. A species can evolve longer hindlimbs only if the

[16]The analysis given in Brown (2014) can be seen as dealing with this causal structure.

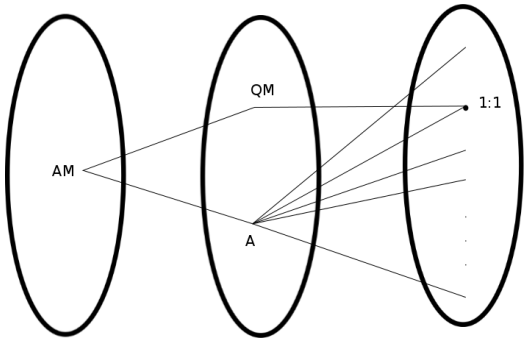

Figure 12: The ancestral money population (AM) branches into quadrupedal monkey (QM) and ape (A). The latter state has the capacity to evolve any limb ratio; the former can only keep the 1:1 ratio.

forelimbs grow by the same amount. However, in ancestral ape populations, a proper subset of quadrupedal monkeys, this constraint was relaxed, to allow for different possible ratios. A more formal representation would look something like figure 12.

Once the intermediary state A is realized, which outcome state (limb ratio) actually reached depends on the environment. In an extreme case, if absolutely no information about A's environment is available, all possible outcomes are to be modelled as equiprobable. This means that once A occurs, it no longer matters for the outcome what preceded that state. The causal network emanating from A constitutes a symmetrical causal tree. To the extent some outcomes can be privileged over others, symmetry is only restored to a certain degree (see next section). In either case, the state A can be thought of as a 'flexible' state: it partially restores symmetry, limited to a subset of the whole causal networked. Thus, while global symmetry once broken cannot ever be restored, global symmetry can be *locally* restored (to a certain degree).

V. Degree of path-dependence

No account of path-dependence can be considered complete without giving some criterion of how history matters more in some processes than in others. We will focus only on how to quantify path-dependence according to how *informationally relevant* the past is for the outcomes. As already mentioned, a possible alternative way to measure path-dependence could be by quantifying how much an outcome changes if past states are changed. This would require the introduction of a separate metric (presumably dependent on explanatory interests) of what it means for outcomes to be close or dis-

tant, with associated problems (see section II). Instead, the focus will be on quantifying the degree of information given by the past in such a way that is consistent with the account of path-dependence presented thusfar.

1. Prediction and Retrodiction

In this approach, path-dependence is closely related to predictability and retrodictability in the following sense. An outcome is more predictable if a past state contains more information about which outcome will occur. Likewise, the past is retrodictable from the present if the outcome contains information about which causal path had been followed.

However, path-dependence precludes both perfect predictability and perfect unpredictability. Recall how a convergent network is perfectly predictable but path-independent, and a maximally divergent network is unpredictable but is also path-independent. In deciding then whether or not a network is path-dependent , it is thus irrelevant whether the outcome can or cannot be predicted from a past state.

The same point can be made about retrodictability. Thus it may be possible to know with fair certainty what causal path the system has followed, but for the network still to be symmetrical and hence path-independent. In Figure 9, we can know with fair certainty, given o_1, that the system passed through s_1, even though passing through s_1 did not affect the probability of o_1. Retrodictability is possible despite path-independence towards o_1. Conversely, the outcome state may not contain any information about which causal path was followed, and yet the network can be path-dependent. This is the case in Figure 10, where both paths leading to o_2 are equiprobable, but where the choice between s_1 and s_2 affects the probability of the outcome.

The relation between predictability (retrodictability) and path-dependence can be made more precise by observing that the amount of information the past *contains* about the present is not relevant for path-dependence, but rather that the past *contributes* to predictability. Thus, in a network converging on a single outcome, the outcome is perfectly predictable regardless of whether the precise past state is known. However, knowing the past does not *contribute* any information not already contained by the structure of the causal network. Neither does it matter how much information the present contains about the past, but only how much the present affects retrodictability. In Figure 10, knowing which of the two intermediary states is reached helps to predict which of the three outcomes is likely to occur, whereas knowing which outcome occurred affects retrodictability.

One may wonder here if contribution to predictability and contribution to retrodictability are equivalent. If they were not equivalent, one would need

to distinguish between two measures of path-dependence: a forward-oriented and a past-oriented measure. However, it is straightforward to show that they are equivalent.

Assume the past does not affect predictability, then the probability of an outcome conditional on an earlier state is simply the unconditional probability: $P(o_j|s) = P(o_j)$. Thus, in a network where the past does not contribute to predictability, the conditional probability of an outcome is equal to the unconditional probability. Similarly, it is the contribution of the present to the retrodictability of the past that matters. It does not matter when $P(s|o_j) = P(s)$ for every previous state s of a given outcome o_j. We would want to show that if $P(o_j|s) = P(o_j)$ for every outcome o_j and intermediate state s, then $P(s|o_j) = P(s)$ (and vice versa).

From Bayes' rule,

$$P(s|o_j) = \frac{P(o_j|s)P(s)}{P(o_j)}$$

and the desired result follows from the assumption that the past does not affect predictability. Thus it is impossible for the past to affect predictability without the present affecting retrodictability, and vice versa.

2. Mutual information

Predictability is the lack of uncertainty of what the outcome state will be. Thus maximal unpredictability corresponds to a uniform probability distribution over the possible outcomes; maximal predictability assigns probability 1 to a single outcome and zero to the rest. In this way the **conditional entropy** of O given s,

$$H_s(O) = -\sum_o P_s(o) \log P_s(o) = -\sum_o P(o|s) \log P(o|s),$$

is a good measure for how predictable the outcomes seem from the perspective of intermediary state s. It has a number of desirable properties: it is maximal for a uniform distribution, and zero when one of the outcomes is certain. A different conditional entropy, of O given S is obtained by taking the weighted average over the states in S:

$$H(O|S) = \sum_s P(s) H_s(O).$$

The extent to which knowing past states S reduces uncertainty — the quantity, we have argued, relevant to path-dependence as symmetry breaking

— is measured by the **mutual information** between the outcome states O and the set of past states S at some instant t:

$$I(O;S) = \sum_{o,s} p(o,s) \log \frac{p(o,s)}{p(o)p(s)} \tag{1}$$

Note that this formulation of mutual information is a measure of path-dependence localized to a particular instant in the causal network. Analogous measures can be formulated for the other notions of symmetry (both local and global); however, the measure 1 is sufficient to extract the philosophically interesting properties.

Mutual information is consistent with the symmetry account of path-dependence in many different respects. First, mutual information is nonnegative $I(O;S) \geq 0$, and zero if and only if the causal network is symmetric at $s \in S$. This can be seen as follows. If the network is symmetric at s, then for any given outcome state o and $s^* \in S$: $p(o|s) = p(o|s^*)$. From this and Theorem 2 can be deduced that these conditional probabilities are equal for all ancestor nodes, including any of the initial states s_0: $p(o|s) = p(o|s_0) = p(o)$. In this case $p(o,s) = p(o|s)p(s) = p(o)p(s)$ and hence

$$\begin{aligned} I(O;S) &= \sum_{o,s} p(o)p(s) \log 1 \\ &= 0 \end{aligned}$$

The mutual information is zero. The opposite also holds true: if mutual information is zero between O and S, then $p(o,s) = p(o)p(s)$ for every $s \in S$.[17] This implies symmetry.

Second, the claim that path-dependence is to be measured by information-contribution rather than information-content is underlined by the relation between mutual information and Shannon entropy.[18] Mutual information represents the information *gain* represented by an intermediate state:

$$I(O;S) = H(O) - H(O|S) \tag{2}$$

Thus, the degree of path-dependence is measured by the reduction in the uncertainty of the outcome states when information about a later intermediary state S is integrated. Path-independence arises when there is no change in entropy content.

[17] For the derivation, see *e.g.* Cover and Thomas 1991, Ch. 2

[18] See Cover and Thomas 1991.

This suggests another way of viewing this aspect of path-dependence, in terms of the divergence of probability distributions. Mutual information can be expressed as the degree by which the unconditional $p(O)$ and the conditional distribution $p(O|S)$ *diverge*.[19] When there is no divergence, $p(O|S) = p(O)$ and the outcome states are independent of the intermediary states S. Thus also in this respect, mutual information seems to be a natural operationalization of the symmetry formulation of path-dependence.

Third, mutual information is symmetric, *i.e.* $I(O;S) = I(S;O)$. This means that the present is relevant for the past in exactly the same way that the past is relevant for the present. This allows the previous arguments about the relation between path-dependence and predictability to be represented more formally. Here follows the case for predictability; identical reasoning can be applied to retrodictability (where $H_o(S)$ is the relevant measure for retrodictability). Perfect unpredictability means that the conditional entropy of the outcome states O is maximal, at any given set of intermediary states S. This means that the unconditional entropy $H(O)$, which, in our framework, is the conditional entropy given the initial states S_0, is also maximal. Hence the mutual information $I(O;S)$ is zero, implying path-independence. Perfect predictability implies that the unconditional entropy $H_s(O)$ is zero at every S; hence $I(O;S)$ is likewise zero.

This operationalization allows for information-theoretic analyses of path-dependence. Two interesting lines of inquiry for further research can be pointed to. A first concerns how mutual information changes as the grain of analysis changes. Thus, in the introduction we outlined how the path-dependence of a process depended on how both the initial states and the outcome states were described. The same process could be described as path-dependent and as path-independent. We already showed how fine-graining and coarse-graining had an effect on the convergence and divergence of a network; hence, one would expect the fine-graining of the outcomes to increase mutual information and thus path-dependence. With this in mind, we can conjecture that describing the outcome states at a more detailed grain of analysis increases the degree of path-dependence:

In a given causal network (O, I, R), if $O = \{o_1, o_2, \ldots, o_n\}$ is fine-

[19]The technical expression is that mutual information is the expectation, given S, of the Kullback-Leibler divergence between the distribution $p(O)$ and the conditional distribution $p(O|S)$:

$$I(O;S) = \mathbb{E}_S \left[D_{KL}(p(o|s)||p(o)) \right].$$

. This is simply a quantitative expression of the how much the conditional probability distribution is expected to diverge from the unconditional distribution, 'from the perspective' of some time in the past.

grained to $O' = \{o_{11}, \ldots, o_{1k_1}, o_{21}, \ldots, o_{2k_2}, \ldots, o_{n1}, \ldots, o_{nk_n}\}$, then $I(O'; S) \geq I(O; S)$.

The second line of inquiry would be to investigate how mutual information changes over time, and how it is affected by symmetry breaking.[20] For example, an interesting consequence of the nonnegativity of mutual information is that, through equation (2), the conditional entropy at some intermediary state is never greater than the unconditional entropy: $H(O) \geq H(O|S)$. The entropy $H(O)$ can be thought of as the uncertainty on the distribution of outcome states without knowing anything about the past (*i.e.* the difficulty in reconstructing the outcome distribution). In this way the inequality means that knowing some intermediary state will never increase the uncertainty over the outcome states.

What is of interest is how the conditional entropy evolves over time $H(O|S)$. While a analysis in full generality is beyond the scope of this paper, two simple cases can be mentioned. The first concerns the case where a network remains symmetric until some intermediate set of states S, after which the symmetry is broken. From (2) follows that the mutual information is zero at all intermediary states S^* before S, and from (2), this means that $H(O) = H(O|S^*)$. Thus the conditional entropy remains constant until the breaking of the symmetry, after which it monotonically decreases. This is the same result, derived by different means, as in theorem 2.

A second simple case is when the network is a causal tree. Here $H(O) = H(O|S_0)$ (since there is only one initial state), and each branching even creates a sub-tree. Hence $H(O) = H(O|S_0) \geq H(O|S_1) \geq H(O|S_2) \ldots$, and conditional entropy monotonically decreases over time. In a branching tree, later states always contain more information about the outcome than the initial states do.

Such a result may seem counterintuitive at first. Cannot a network start off with a bias towards some outcomes, and then evolve towards a uniform distribution, such as in Figure 13? Would this not increase the conditional entropy? The answer is that the network does not evolve towards a uniform distribution over *all* possible outcomes. The evolution towards uniformity is outweighed by the fact that any branch will have some inaccessible outcomes. Thus, while s_0 branches out to four different outcomes, s_1 and s_2 branch out to only two different

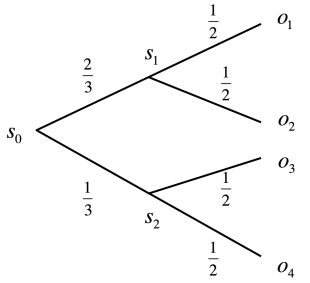

Figure 13: Decrease of entropy, despite uniformity.

[20] See Sober and Steel (2011) for a related analysis of entropy change in Markov models. Since causal networks are Markovian, many of their results would also be applicable here.

outcomes. The entropy of four equiprobable outcomes is $\log 4$, whereas the entropy of two equiprobable outcomes is $\log 2$. In this case, $H(O|S) = \log 2$ and $H(O|S_0) = \frac{2}{3}\log 3 + \frac{1}{3}\log 6 > \log 2$. In this way, entropy also decreases here over time.

VI. Discussion and conclusion

This article has attempted to give some more formal basis to the concept of path-dependence, and has argued that this is best done by means of symmetry considerations in the framework of causal networks. The two alternative frameworks, attractor landscapes and causal trees, are not only less general than causal networks, and thus not applicable to a wide range of cases, but are also mere limiting cases of causal networks. In particular, networks tend to reduce to trees when the dimensionality of state space is high (in virtue of less convergence); and to landscapes when the dimensionality is low (in virtue of more convergence).

Within the causal network framework, symmetry considerations allow for a both technically and intuitively powerful way of describing path-dependence. Symmetry comes in different scopes — some more global, others more local — and path-dependence in its most general form arises when global symmetry is broken. Local symmetries can be broken for specific states, or at specific times, and in this way path-dependence is something that can *emerge* at a certain point in time. Other interesting phenomena include privileged states that preserve and states that restore local symmetries (weak global attractors and flexible states).

The degree to which global symmetry is broken or preserved can be quantified by means of mutual information. This measure, which quantifies how much information one variable contributes about another, is perhaps not the only possible way to quantify path-dependence and historicity; however, it is conceptually continuous with the qualitative account presented in this paper, and has the added advantage of opening up an information-theoretic perspective on path-dependence.

As a final discussion, I would like to suggest some broader philosophical themes implicit in the symmetry account, in particular, the way in which historicity is intertwined with causal structure and complexity. The symmetry formulation of path-dependence states that, as long as a network remains symmetrical, history does not matter for the eventual outcome. One way in which this can be rephrased is that, in a symmetrical network, it could very well be that time did not pass at all, or passed very slowly. The *durational*

aspect of time makes no difference. Prior to the symmetry breaking, the network can be extended or compressed, added to or subtracted from arbitrarily, without this making any difference for whatever outcome would eventuate. If we allow the durational aspect of time to be represented by the quantity of causal transitions in a network, time only emerges as a relevant physical quantity when the causal symmetry is broken.

Such considerations offer a novel perspective to Curie's general claim, "It is dissymmetry that creates the phenomenon"[21]. While originally formulated in a different context and with a different notion of symmetry in mind (see Earman 2004 or Brading and Castellani 2007 for a discussion of this), in the context of the present paper Curie's claim becomes relevant if one understands 'phenomenon' as 'event of historical significance'. Events are represented in causal networks by the nodes, from which the vertices branch out; for these events to count as 'phenomena', they must lead to the breaking of the symmetry. An analogue in modern physics would be the collapse of the wave function in the standard interpretation of quantum theory: the event of observing the spin of the electron causes the possible causal paths to branch out, and only one path to be realized (either spin up or spin down).[22] At some deep level, observability seems to be connected to the breaking of symmetry.

What is a curious fact is that such causal symmetries are *continually* broken in many dynamics in the special sciences (see Longo and Montévil 2012 for the formulation of a similar idea). History matters at every instant. For example, in biological evolution, with each new evolutionary development, new constraints are set in place, and the set of possible outcomes is made smaller. Exceptions are, of course, possible, but seem to be limited to the global preservation of local symmetries (convergent evolution), or the local restoration of global symmetry (evolvability). In any case, such exceptions are rare occurrences in the space of biological possibility; as a rule, most evolutionary processes are historical and path-dependent.

This observation about biological evolution has been long recognized (see Beatty 1995); however, a similar conclusion would seem to be applicable to special sciences in general. This underlines the fact that explanations of the deductive kind (*e.g.* D-N explanations) are particularly inadequate when it comes to the sciences of complex systems, and suggests that path-dependence is more fundamental to scientific explanation than is currently acknowledged by the philosophical literature.

[21] "C'est la dissymétrie qui crée le phénomène." (Curie, 1894, 400).

[22] Or, if one adheres to the Everettian interpretation, both these paths are realized, but in different parallel universes.

Why this should be — why some processes are much more historical than others — is a deep question that can only be posed in this context. If one considers only the contrast between biology and physics, one factor that would seem relevant is complexity. In statistical physics, an individual entity might have only six degrees of freedom (three for position, three for momentum); in ecology, an individual biological organism has an intractable number of degrees of freedom. On the microevolutionary scale it is feasible to abstract away from this multitude of variables and focus only on on a very limited number of causally relevant variables (*i.e.*, traits). This is one reason why population genetics can be so elegantly mathematicized.[23] By contrast, on the macroevolutionary scale, such a limitation of state space does not seem possible, and state space seems to be necessarily high-dimensional (because potentially every trait can be relevant as environments change). This fact would go part of the way to explaining why the vast majority of macroevolutionary processes seem to be unavoidably historical and path-dependent.

Thus the question as to why historicity emerges in complex systems seems to be intimately related to the question of why simplicity emerges in complex systems (Strevens 2006). Some complex systems allow for some degree of prediction, either through laws or through numerical simulations. Others do not, and for these systems narrative, historical explanations seem to be unavoidable, and perhaps are even optimally explanatory.

[23]Whether the mathematization corresponds to causal reality is a different question. This is related to the debate whether drift and natural selection actually pick out causal forces in reality, or are just a statistical abstraction from the high-dimensional state space (see *e.g.* Matthen 2009 for the relation between abstraction and natural selection).

Appendix

Theorem 1. *A coarse-graining of the explanandum makes an explanation increasingly convergent and a coarse-graining of the explanans makes an explanantion increasingly divergent.* s

Proof. We will prove it for an explanation that is purely parallel, thus neither convergent nor divergent. The generalization for a random explanation holds analogously.

Assume a deterministic explanation (O, I, f), so that f is a bijection $f : I \to O$. Define an equivalence relation $\sim$ on O such that $o_1 \sim o_2$ iff $o_1, o_2 \in A$ for some A (dependent on theoretical interests) with $\#A > 1$. Because f is a bijection there exists a uniquely defined $B \in I$ such that $f(B) = A$. Call B the 'basin' and A the 'attractor' of f on I.

Then O/A represents a coarse-graining of the explanandum and I/B a coarse-graining of the explanans. So define an associated function $R_c : I \to \#O/A : i \mapsto f(i)$ and relation $R_d \in I/B \times O = (f^{-1}(o), o)|o \in O$. Because f is a bijection, $\#I = \#O > \#O/A$ and $\#O = \#I > \#I/B$, and hence R_c will be a non-injective surjection, and R_d a non-function. Hence the number convergent structures has increased in explanation $(O/A, I, R_c)$, and the number of divergent structures has increased in $(O, I/B, R_d)$.

Theorem 2. *Let (O, I, R) be symmetrical at some instant in time. Then (O, I, R) is symmetric at all prior instants.*

Proof. Assume (O, I, R) is symmetric at time t, corresponding to the set of intermediate states S. Let S' represent some earlier generation of states. From the local symmetry of (O, I, R) at S we can deduce that $P(o|s^*) = p \in [0, 1]$ for all $s^* \in S$.

Take a random predecessor state $s' \in S'$. Assume it branches out to a number of states $s^* \in S$. Then

$$\begin{aligned} P(o|s') &= \sum_{s^*} P(o|s^*)P(s^*|s') \\ &= p \sum_{s^*} P(s^*|s') \\ &= p \end{aligned}$$

since the sum of the probabilities of all paths leaving s' is 1. Thus the network is symmetric at S'.

This also means that the bias p towards outcome o is preserved as long as the network remains symmetric.

VII. Bibliography

References

Arthur, Brian W. (1994). *Increasing Returns and Path Dependence in the Economy.* University of Michigan Press.

Bassanini, Andrea, and Giovanni Dosi. (1999). "When and How Chance and Human Will Can Twist the Arms of Clio." LEM Working Paper series 05, SantAnna School of Advanced Studies, Pisa.

Beatty, John. (1995). The Evolutionary Contingency Thesis. In G. Wolters, and J.G. Lennox (Eds.), *Concepts, Theories and Rationality in the Biological Sciences* (pp. 45-81). Pittsburgh: University of Pittsburgh Press.

—. (2006). Replaying Life's Tape. *Journal of Philosophy*,103:336-362.

Beatty, John, and Eric Desjardins. (2009). Natural Selection and History. *Biology and Philosophy*, 24:231-246.

Brading, Katherine and Castellani, Elena (Eds.). (2003). *Symmetries in Physics: Philosophical Reflections.* Cambridge: Cambridge University Press

—. (2007). Symmetries and Invariances in Classical Physics. In J. Butterfield and J. Earman (eds.), *Handbook of the Philosophy of Science. Philosophy of Physics* (pp. 13311367). Amsterdam: North Holland, Elsevier.

Cover, Thomas M. and Thomas, Joy A. (1991). *Elements of Information Theory.* John Wiley & Sons.

Curie, Pierre. (1894). Sur la symétrie dans les phénomènes physiques, symétrie dun champ électrique et dun champ magnétique. *Journal de Physique Théorique et Appliquée*, 3(1): 393-415.

David, Paul A. (1985). Clio and the economics of QWERTY. *American Economic Review*, 75: 332337.

Desjardins, Eric. (2011a). Historicity and Experimental Evolution. *Biology and Philosophy*, 26: 339-364.

—. (2011b). Reflections on Path Dependence and Irreversibility: Lessons from Evolutionary Biology. *Philosophy of Science*, 78:724-738.

—. (2015). Historicity and ecological restoration. *Biology and Philosophy*, 30:77-98.

Ereshefsky, Marc. (2012). Homology Thinking. *Biology and Philosophy*, 27:381-400.

Gavrilets, Sergey. (2004). *Fitness landscapes and the origin of species*. Princeton: Princeton University Press.

Gould, Stephen J. (1989). *Wonderful Life: The Burgess Shale and the Nature of History*. W. W. Norton & Co.

Kaplan, Jonathan. (2008). The end of the adaptive landscape metaphor? *Biology and Philosophy*, 23:625-638.

Longo, Giuseppe, and Montévil, Ma el. 2011. From physics to biology by extending criticality and symmetry breakings. *Progress in Biophysics and Molecular Biology*. 106(2): 340-347.

MacKay, David J.C. (2003). Information Theory, Inference, and Learning Algorithms. Cambridge University Press.

Matthen, Mohan. (2009). Drift and 'Statistically Abstractive Explanation'. *Philosophy of Science*, 76: 464-487.

Pierson, Paul. (2004). *Politics in time: history, institutions, and social analysis*. Princeton:Princeton University Press.

Pigliucci, Massimo, and Kaplan, Jonatan. (2006). *Making Sense of Evolution*. University of Chicago press.

Plutynski, A (2008) The rise and fall of the adaptive landscape? Biology and Philosophy, 23:605-623.

Ruse, Michael. (1996). Are Pictures Really Necessary? The Case of Sewall Wrights 'Adaptive Landscapes'. In Baigrie B.S. (ed.), *Picturing Knowledge: Historical and Philosophical Problems Concerning the Use of Art in Science*, 303-337. Toronto: University of Toronto Press.

Skipper, Robert A. (2004). The heuristic role of Sewall Wrights 1932 adaptive landscape diagram. *Philosophy of Science*. 71: 1176-1188.

Sober, Eliott. (1983). Equilibrium Explanation. *Philosophical Studies*, 43:201-210.

—. (1988). *Reconstructing the past: parsimony, evolution, and inference*. Cambridge, MA: MIT Press.

Sober, Eliott, and Steel, Mike. (2011). Entropy increase and information loss in Markov models of evolution. *Biology and Philosophy*, 26:223-250.

Strevens, Michael. (2006). *Bigger than Chaos: Understanding Complexity through Probability.* Cambridge, MA: Harvard University Press.

Velasco, Joel D., and Sober, Eliott. 2010. Testing for treeness: lateral gene transfer, phylogenetic inference, and model selection. *Biology and Philosopy*, 25:675-687.

Wilkins, Jon F., and Peter Godfrey-Smith. (2009). Adaptationism and the adaptive landscape. *Biology and Philosophy*, 24:199-214.

Young, Nathan M. and Hallgrímsson, Benedikt. (2005). Serial Homology and the Evolution of Mammalian Limb Covariation Structure. *Evolution*, 59(12):2691-2704.